ABSTRACT OF THE DISCLOSURE
LAYERED LOCAL CACHE WITH LOWER LEVEL CACHE OPTIMIZING ALLOCATION MECHANISM

A method of improving memory access for a computer
system, by sending load requests to a lower level storage
subsystem along with associated information pertaining to
intended use of the requested information by the requesting
processor, without using a high level load queue. Returning
the requested information to the processor along with the
associated use information allows the information to be
placed immediately without using reload buffers. A register
load bus separate from the cache load bus (and having a
smaller granularity) is used to return the information. An
upper level (L1) cache may then be imprecisely reloaded (the
upper level cache can also be imprecisely reloaded with
store instructions). The lower level (L2) cache can monitor
L1 and L2 cache activity, which can be used to select a
victim cache block in the L1 cache (based on the additional
L2 information), or to select a victim cache block in the L2
cache (based on the additional L1 information). L2 control
of the L1 directory also allows certain snoop requests to be
resolved without waiting for L1 acknowledgement. The
invention can be applied to, e.g., instruction, operand data
and translation caches.